CDC’s Web Thesaurus: a Vocabulary Framework for Public Health

Mamie J. Bell[1], Kathy Lesh[2], Shellie Kolavic Gray[3]

[1]Centers for Disease Control and Prevention, Atlanta, GA; [2]The KEVRIC Company, Inc., Silver Spring, MD; [3]The KEVRIC Company, Inc., Atlanta, GA

Rationale for Work

- No public health terminology system exists
- A discrete vocabulary is needed to provide:
  - A hierarchy of concepts specifically organized for public health
  - An application-generic vocabulary (not for literature tagging or medical coding)
  - Robust ontological relationships among concepts (beyond Broader Term and Related Term)
  - A synonymy reflective of public health, not biomedicine (eg, lay terms & acronyms)
  - Rapid concept additions and distributions to the Public Health Workforce
  - Rapid retiring of outdated concepts

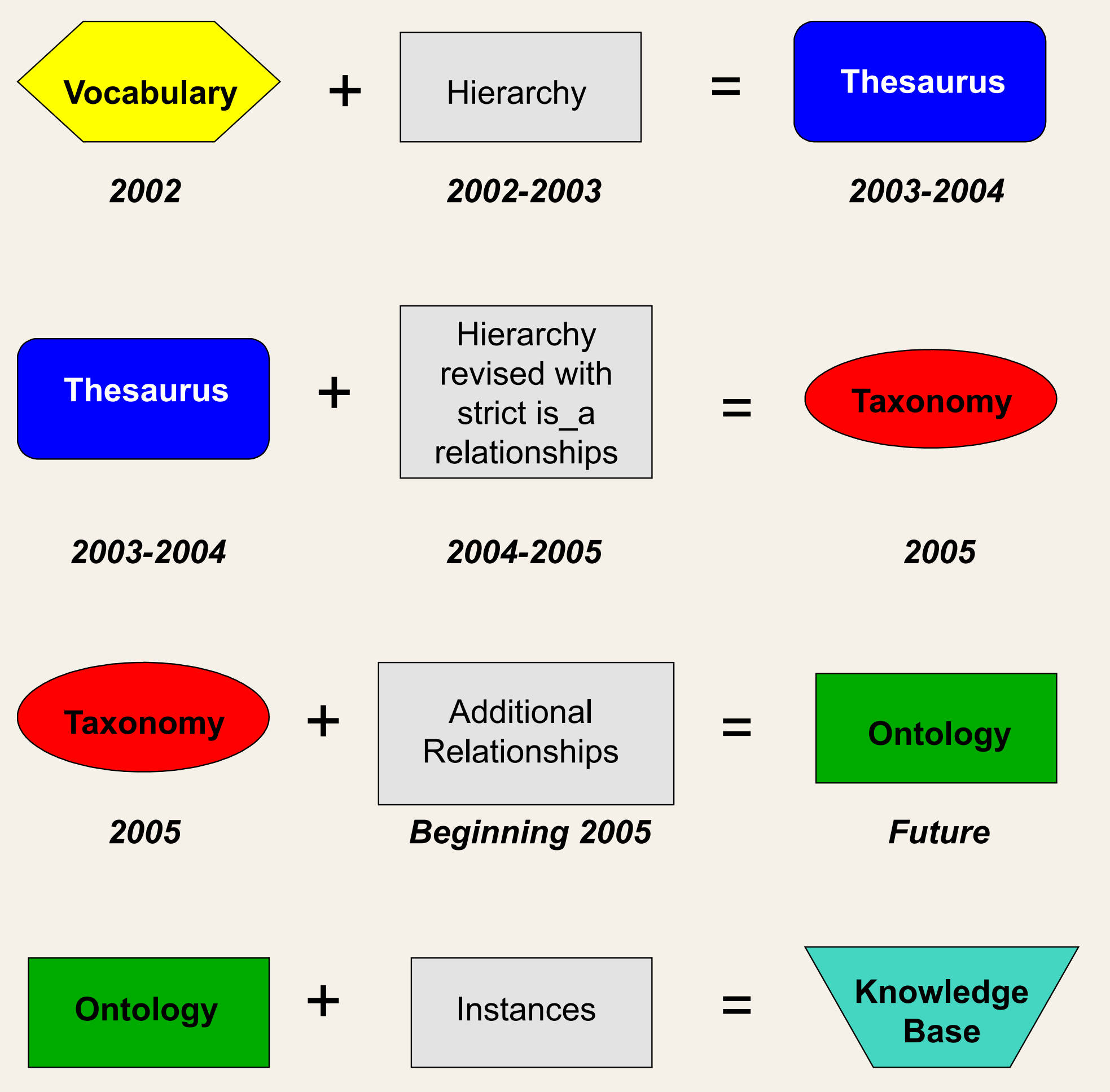
Initial Research & Build of the Controlled Health Thesaurus (CHT)

- Public Health Concepts identified by
  - Reviewing existing CDC vocabularies
  - Extracting terms from the CDC website and search logs
- Matched these public health terms to the UMLS to select vocabularies that offered the widest coverage
  - Matching revealed a Public Health concept “Gap” of up to 50%
- Imported concepts with their hierarchical structures from these standard vocabularies for the base Thesaurus:
  - Medical Subject Headings (MeSH)
  - Computerized Retrieval of Information on Scientific Projects (CRISP)
  - Alcohol and Other Drug Thesaurus (AOD)

Definitions

- **Concept**- a notion, an idea, a unit of thought
- **Term** - a word or words corresponding to one or more concepts, a linguistic label
- **Synonym** - a term which is an acceptable alternative to the preferred term as a way of expressing a concept
- **Is_a**- a relationship identified in a taxonomy: concept A is a child of concept B if every instance of A is also an instance of B. For example, apple is a child of fruit; therefore every apple is a fruit.

CHT Evolution



Value of Changes

- **Hierarchy** - infers some relationship between concepts
- **Is_a Hierarchy** – clearly denotes relationship between parent and child concepts; allows for an implicit definition of the child based on inheritance of the parent’s characteristics
- **Additional Relationships** – establishes relationships between concepts beyond that of parent/child to explicitly state how concepts can be used together
- **Instance** – provides specific pieces of data for concepts which transforms information about a domain into knowledge (i.e. VIN # JH4NA1152MT001365 is an instance of a vehicle)

Thesaurus ➡ Taxonomy

- CHT initially built as a thesaurus following **ANSI/NISO Z39.1993** – Guidelines for the Construction, Format, and Management of Monolingual Thesauri
  - Children RELATED to parent, but not necessarily is_a
- CHT now a taxonomy moving toward ontology following **ISO TC 17117** – Controlled health vocabularies – Vocabulary structure and high level indicators
  - Taxonomic structure (true is_a)
- **ISO 11179** – Information technology – Specification and standardization of data elements (a.k.a. metadata), also being followed to ensure standardization across CDC

Other Standards Used for Reference in Taxonomic Expansion of the CHT

- **For anatomic concepts**
  - Foundational Model of Anatomy (FMA)
- **For organisms**
  - NCBI Taxonomy
  - ICTV Taxonomy and Index to Virus Classification and Nomenclature
  - Integrated Taxonomic Information System
- **For chemicals and drugs**
  - NDF-RT
- **For all areas**
  - NCI Metathesaurus

Criteria for Concept Inclusion

- Each imported term and synonym was reviewed for the following before being confirmed for inclusion into the CHT:
  - Hits on the CDC web site at least 5 or more times
    - Hits were reviewed to ensure SEMANTIC match rather than SYNTACTIC match

  OR

  - Needed as a parent concept for more specific terms
    - <Homeland security related activity> parent for <Counterterrorism activity> and <Terrorism>

Top Level Node Changes

The original top level nodes (imported from MeSH) were revised to meet the public health needs:

New Top Level Nodes

- Activities and procedures
- CDC administrative concepts
- Events
- Findings
- Injuries
- Methods and techniques
- Physical objects
- Population group
- Processes and phenomena
- Properties and attributes
- Substances

Original Top Level Nodes Remaining

- Organisms
- Anatomy
- Diseases and Conditions split:
  - **Diseases**
  - **Injuries**

Original Top Level Nodes Demoted

- Food and Beverages – 2nd level
- Persons – 2nd level
- Behavior – 2nd level
- Biological Sciences – 3rd level
- Health Care – 3rd level
- Information Science – 3rd level
- Chemicals and Drugs split and demoted to 2nd level
- Geographic locations – 2nd level

Original Top Level Nodes Retired

- Anthropology, Education, Sociology and Social Phenomena
- Analytical, Diagnostic and Therapeutic Techniques and Equipment
- Units and Other Authority Lists
- Psychiatry and Psychology

Filling the Public Health “Gap”

- Current standard biomedical vocabularies do not represent the public health terminology needed by CDC
- Initial Research revealed that the largest concept deficits fell into the areas of:
  - Occupational Health and Safety
  - Environmental Health
  - Injury Prevention & Safety
  - Organisms / Organism-related Diseases
  - Bioterrorism and Preparedness Response
- In addition to building out the known gap, CDC is adding concepts needed for emerging infectious diseases and public health events

Example “Gap” Concept: Tsunami

A discrete concept <Tsunami> did not exist in standard biomedical vocabularies on Dec 27, 2004.

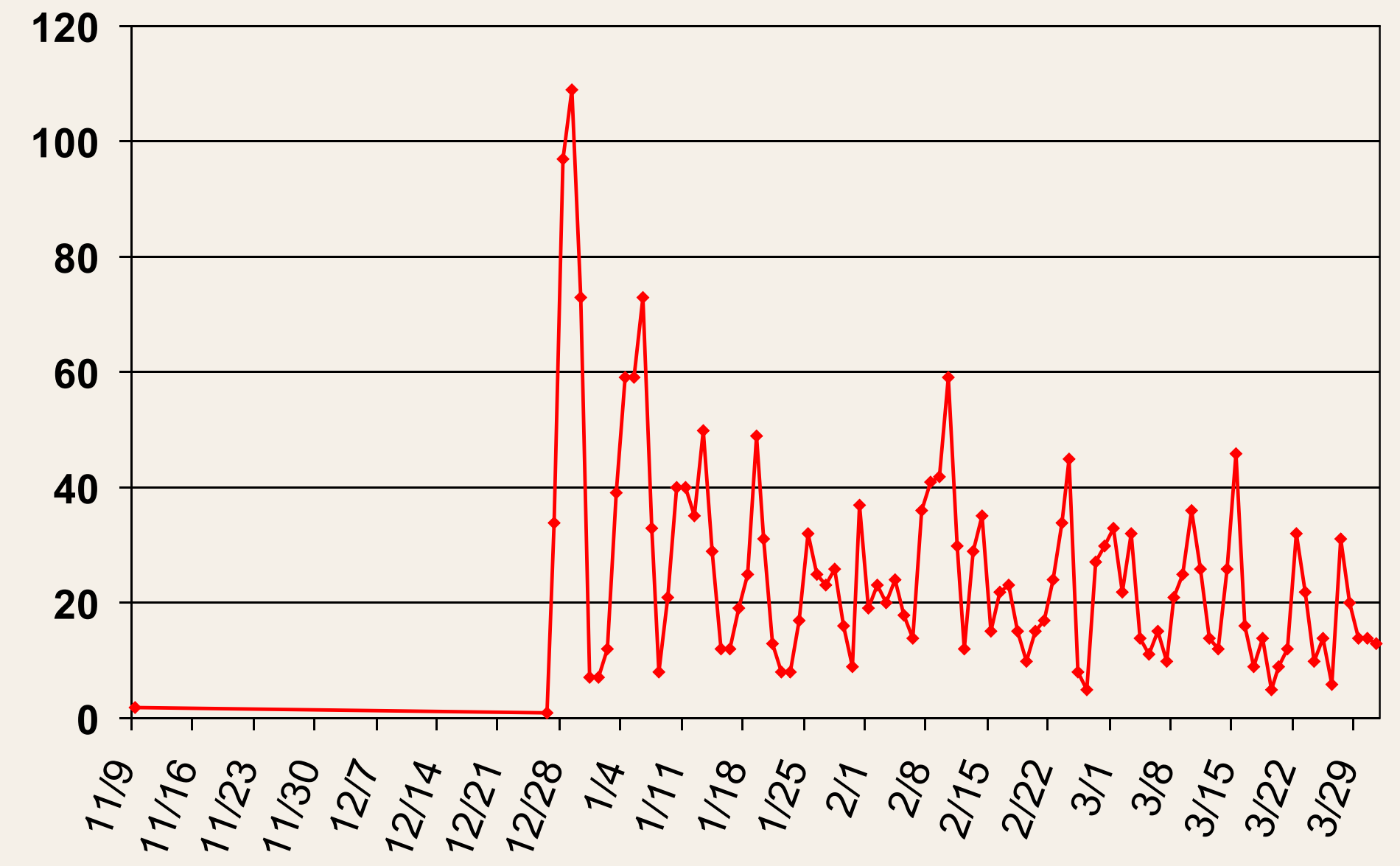
Terminology System	UMLS Metathesaurus Search Results for <tsunami>
ICD-9-CM	<Accident due to tsunami>
SNOMED CT	<Accident caused by tsunami>
MeSH	Not found
Other UMLS Vocabularies	Not found

Concept <Tsunami> Needed

- The public turned to the CDC website for Tsunami-related public health information
- A fully-defined concept with accompanying metadata was needed for tagging CDC web content
  - Preferred term
  - Definition
  - Parents/Children
  - Semantic Type
  - Synonyms
- CDC website hits for <Tsunami> increased dramatically in late December 2004
- CDC responded by adding 100+ pieces of content

CDC Website Searches for Tsunami

November 2004 through March 2005



Next Steps for CHT

- Development
  - Continue to build out the Public Health Gap
    - Leverage CDC subject matter expertise on public health topics
    - Collaborate with CDC Public Health Partners in their domain areas
    - Enable term submission by the Public Health Workforce
- Distribution
  - Draft CHT now available for local use
  - Provide enhanced search/browse and hierarchy view via web browser access
  - Submit concepts to standard vocabularies
  - Submit to aggregator such as the UMLS Metathesaurus

References:

- American National Standards Institute/National Information Standards Organization. ANSI/NISO Z39.19 – Guidelines for the Construction, Format, and Management of Monolingual Thesauri.
- University of Washington. Foundational Model of Anatomy. http://sig.biostr.washington.edu/projects/fm/.
- International Committee on Taxonomy of Viruses (ICTV). ICTV Taxonomy and Index to Virus Classification and Nomenclature. http://www.ncbi.nlm.nih.gov/ICTVdb/Ictv/fr-indv0.htm.
- Integrated Taxonomic Information System (ITIS). http://www.itis.usda.gov/.
- International Organization for Standardization (ISO). ISO TC 17117, Controlled health vocabularies – Vocabulary structure and high level indicators.
- International Organization for Standardization (ISO). ISO/IEC 11179, Information technology – Specification and standardization of data elements.
- National Center for Biotechnology Information (NCBI). NCBI Taxonomy. http://www.ncbi.nlm.nih.gov/entrez/linkout/tutorial/taxtour.html.
- National Cancer Institute (NCI). NCI Thesaurus. http://nciterms.nci.nih.gov/NCIBrowser/Dictionary.do.
- National Drug File Reference Terminology (NDF-RT). Http://nciterms.nci.nih.gov/NCIBrowser/Connect.do?dictionary=VA_NDFRT&bookma.